CopyCat – The Cophylogenetic Analysis Tool User Manual

CopyCat is a software tool written in Java which provides an easy and fast access to cophylogenetic analyses. It incorporates a wrapper for the program ParaFit, which conducts statistical tests for the presence of global congruence between host and parasite phylogenies and for the significance of individual host-parasite associations[1]. CopyCat offers various features, such as the creation of customized host-parasite association data, the reconstruction of host or parasite trees from the NCBI taxonomy, and the computation of several tree statistics. As of April 2007, CopyCat supports Alexandros Stamatakis' programs AxParafit and AxPcoords, which are highly optimized versions of Parafit and DistPCoA, respectively (see section “CopyCat – Available menu bar options”). This manual describes the features specific to CopyCat; regarding the principles of the statistical tests implemented in ParaFit, users are strongly advised to consult Legendre et al. (2002). The literature cited in the references section is also suggested for further reading.

If you use CopyCat, you should cite the accompanying paper:

Meier-Kolthoff J.P., Auch A.F., Huson D.H., Göker M. COPYCAT: Co-phylogenetic Analysis tool. Bioinformatics. 2007. Epub ahead of print

Check list

After installing CopyCat, please make sure that you have verified the following items:

- You have either a Win32-compatible or a Linux-based operating system (currently tested under Windows 2000, Windows XP and Linux x86 32bit with GTK 2.0)
- Your machine is equipped with at least 512 MB of memory. Even though it is not recommended, hardware with less memory can sometimes be used as long as the NCBI taxonomy facilities are not applied.
- The Java 1.5 runtime environment is installed (by entering the command "java -version" in a command console the currently installed version is reported) and the Java Binary must be included in the PATH environment variable (this is done automatically by the Java Installer for Windows). Java 1.5 (a.k.a. JDK5) is available here: http://java.sun.com/javase/downloads/index.jsp
- You have downloaded the current NCBI taxonomy file ("taxdmp.zip") from the following URL: ftp://ftp.ncbi.nih.gov/pub/taxonomy/taxdmp.zip and placed this file in the "input-data" subfolder. It is located in your CopyCat installation folder (e.g., /home/john/CopyCat/input-data/). **As the NCBI taxonomy is updated on a regular basis it is advisable to get the latest version of "taxdmp.zip" from time to time.**

1 Of course, any kind of associations with hosts can be examined in that way, including mutualists. We refer to parasites only just for convenience.

The first start

Once you have started CopyCat for the first time, the following configuration dialogue will appear[2].

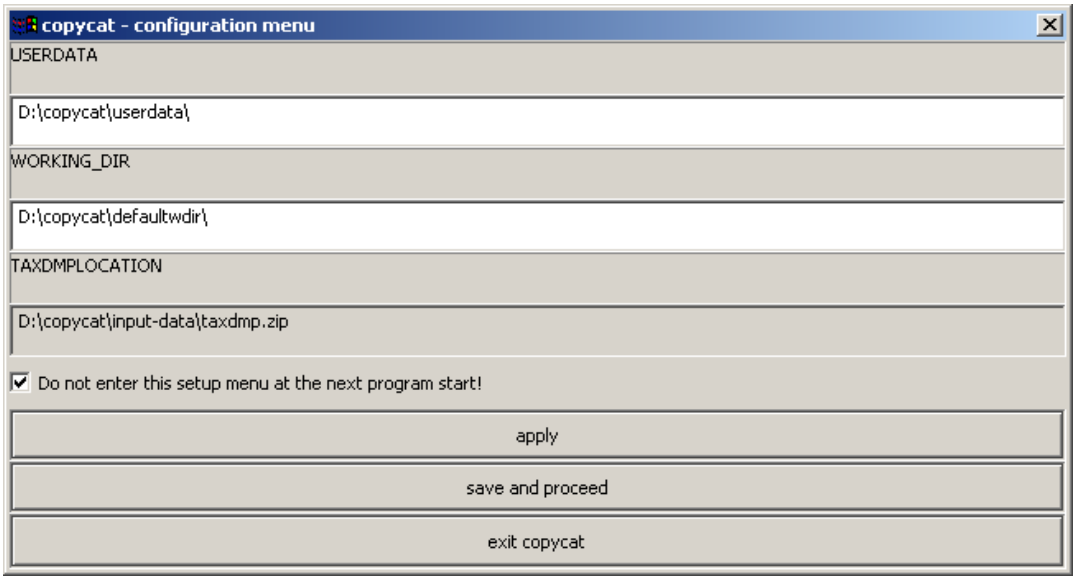


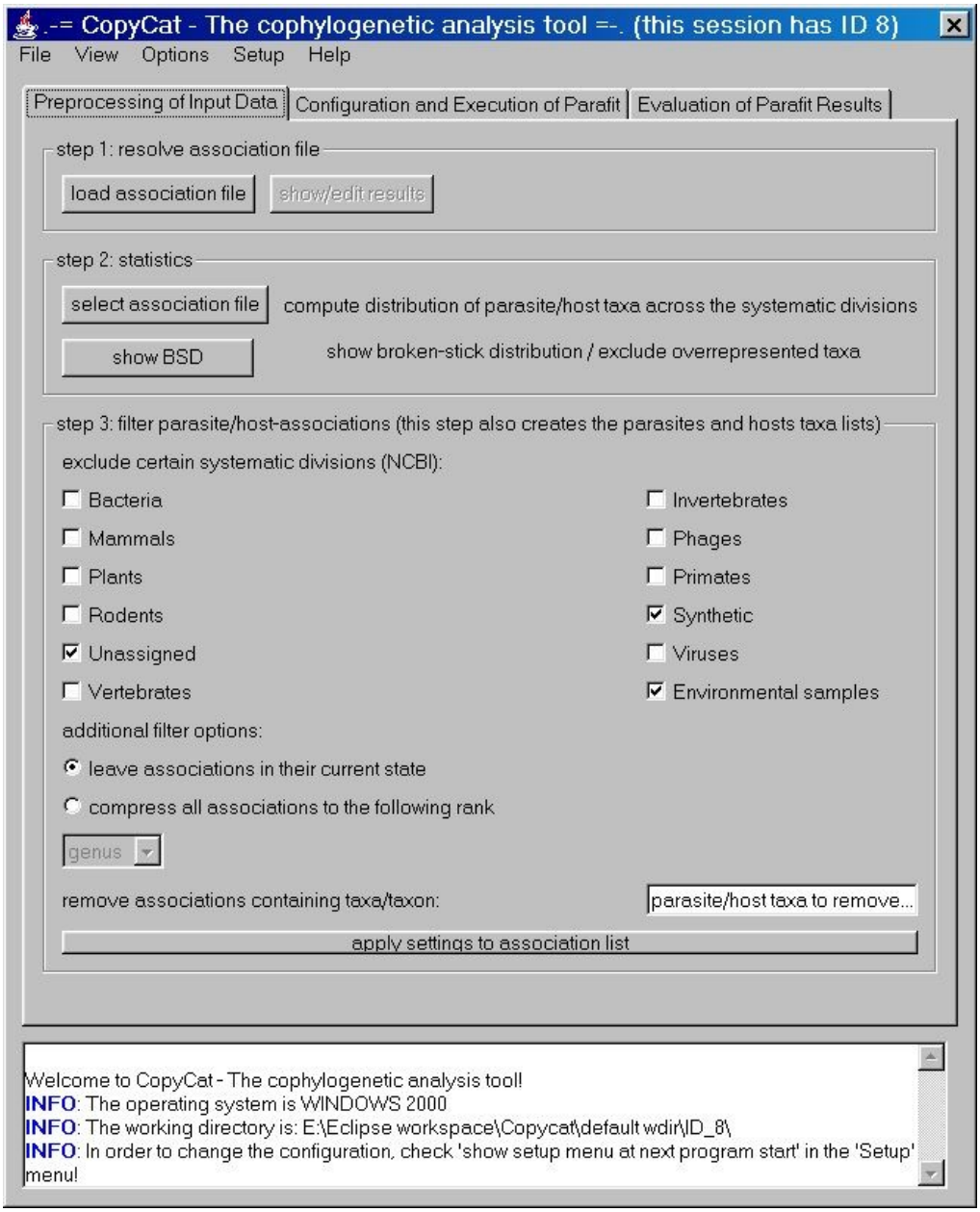
Figure 1: CopyCat's configuration dialog

Here, you might want to choose another working directory ("WORKING_DIR") or another directory for your custom data ("USERDATA"). The latter will be relevant if CopyCat asks for a file to open - then this directory will be prompted first.

To proceed, please click "apply" followed by "save and proceed". By pressing "exit copycat" the whole application will close. If the configuration is not complete the dialogue will be displayed again until everything is well-configured.

Now, the CopyCat application launches. It generates a random session ID (here: ID 9511) and creates a subdirectory for this session within your working directory (e.g., D:\CopyCat\myworkingdir\ID_9511\). All files and results are now going to be stored in that directory. The contents of that folder and thus the results of this session can be examined by selecting the "View" menu item.

2The following screen shots were made using the Windows version of CopyCat - the "look & feel" of the Linux version differs slightly.


.-= CopyCat - The cophylogenetic analysis tool =-. (this session has ID 8)
File View Options Setup Help
Preprocessing of Input Data
Configuration and Execution of Parafit
Evaluation of Parafit Results
step 1: resolve association file
load association file
show/edit results
step 2: statistics
select association file
compute distribution of parasite/host taxa across the systematic divisions
show BSD
show broken-stick distribution / exclude overrepresented taxa
step 3: filter parasite/host-associations (this step also creates the parasites and hosts taxa lists)
exclude certain systematic divisions (NCBI):
Bacteria
Invertebrates
Mammals
Phages
Plants
Primates
Rodents
Synthetic
Unassigned
Viruses
Vertebrates
Environmental samples
additional filter options:
leave associations in their current state
compress all associations to the following rank
genus
remove associations containing taxa/taxon:
parasite/host taxa to remove...
apply settings to association list
Welcome to CopyCat - The cophylogenetic analysis tool!
INFO: The operating system is WINDOWS 2000
INFO: The working directory is: E:\Eclipse workspace\Copycat\default wdir\ID_8\
INFO: In order to change the configuration, check 'show setup menu at next program start' in the 'Setup' menu!
Figure 2: The first tab of CopyCat

CopyCat - The first tab

The first tab (see figure 2) deals with the creation and pre-processing of an association table containing parasite or host associations.

Finding NCBI taxonomy IDs

Given an (unresolved) association list containing parasite or host names, the "resolve association list" item (see figure 3) tries to assign a NCBI taxonomy ID for each organism name. The set of IDs is necessary for inferring the NCBI host or parasite tree.

The user specifies a file (unresolved association table) containing one parasite and one host name per line (tab-separated). Then CopyCat tries to retrieve a NCBI taxonomy ID equivalent to each entry (e.g., the taxonomy ID "9606" for the host name "Homo sapiens").

These results can be watched by clicking the "show/edit results" button. This button is enabled once the results are complete. Figure 3 shows an example of this result window.

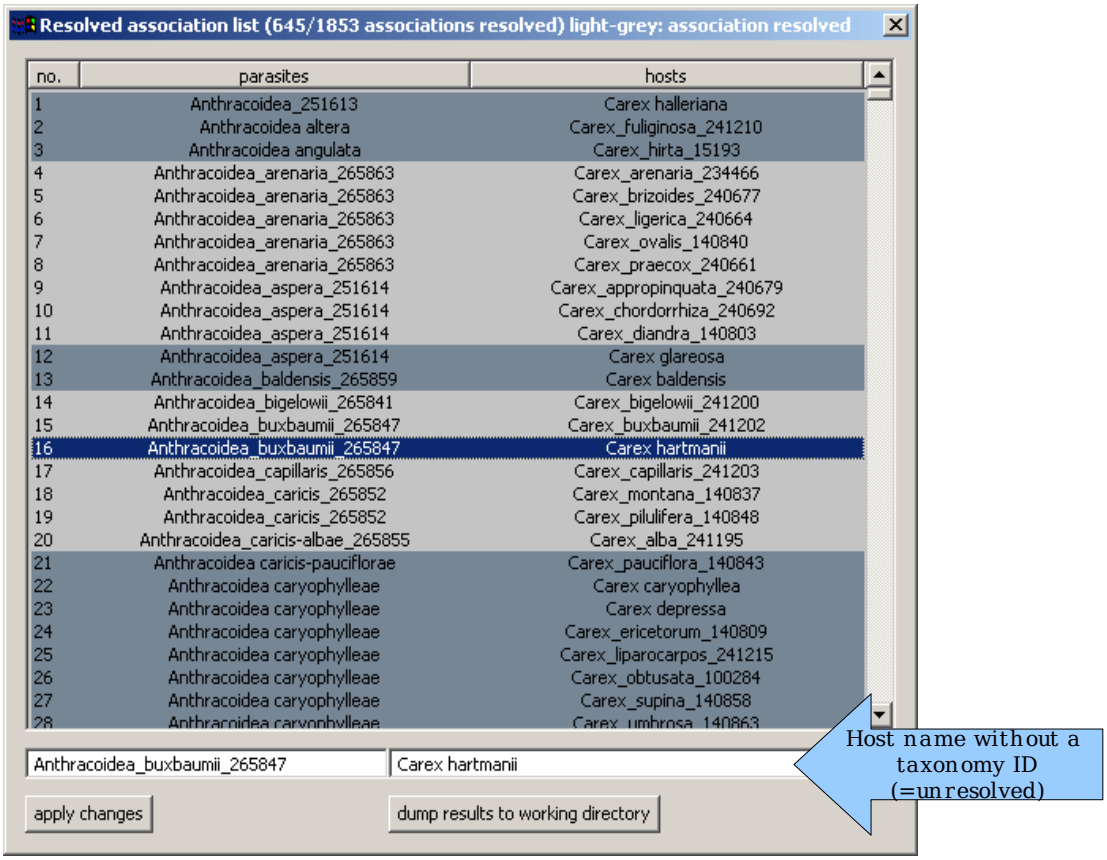


Figure 3: An association table as resolved by CopyCat. The entries coloured in light-grey have for both parasite and host a valid ID assigned as provided by the NCBI Taxonomy Database.

If for both parasite and host name a respective taxon ID is retrieved, this parasite-host-association appears in the final (resolved) association table. The set of all host or parasite taxa contained in that table is used for the reconstruction of the NCBI host or parasite tree, respectively.

If a certain parasite or host name can't be resolved (e.g., because of a misspelling), the user has the possibility to manually enter a proper NCBI taxonomy ID. The format for this is the organism name followed by the taxonomy ID, e.g., "Homo sapiens 9606". The user should ascertain that this taxonomy ID is definitely contained in the NCBI taxonomy database. The better and most consistent way of retrieving a correct ID is as follows: The user corrects the organism name and then applies the changes by pressing the "apply" button. By entering a name available in the NCBI taxonomy, the respective organism field in the association table should then yield the corresponding ID.

Finally, all resolvable associations are written to a file after having selected the "dump results to working directory" button/option. Along the new association file, two files containing only the parasites and, respectively, only

the hosts, are written to the same directory.

Statistics

Characteristics of the association table

The user hits the "select association file"-button and then specifies an association table (AT; refer to first step). The AT is read and the following features are reported to the message window:

- The distribution of the taxonomical ranks within the set of hosts and the set of parasites respectively (in accordance to the NCBI taxonomy)
- The affiliation of each parasite and host taxon, respectively, to one of the following divisions:

  division 0: Archaea

  division 1: Bacteria

  division 2: Eukaryota

  division 3: Viroids

  division 4: Viruses

  division 5: Other

  division 6: Unclassified

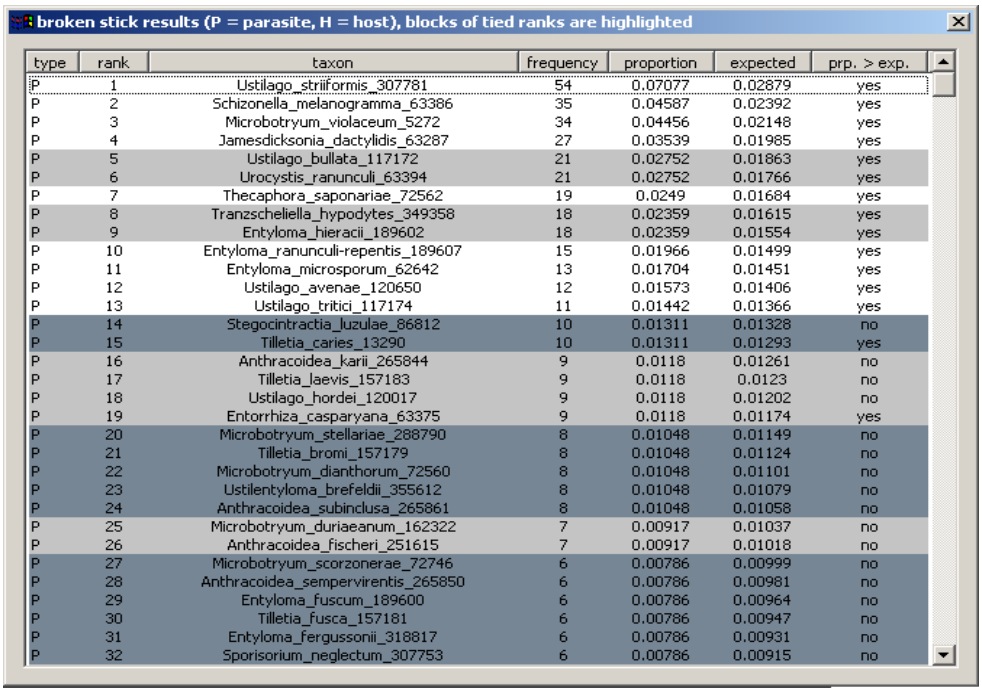
  division 7: taxon ID not found in NCBI taxonomy
- The number of associations contained in the AT
- The number of different parasite taxa in AT
- The number of different host taxa in AT
- The estimated size of the association matrix drawn from the AT

N.B.: As a rule, a phylogenetic tree derived from character data (e.g., molecular sequences) contains more information than a taxonomy of the same taxa since the topology of the latter is usually much less resolved. However, if a cophylogenetic study is based on specific marker sequences such as 16S-rRNA or ITS, it is limited to species for which there is a common marker gene available. Even though the number of single-locus or even genome sequences is steadily increasing, we presume that NCBI taxonomic data are available for many more taxa than are such orthologous genetic data. (We guess this rule also holds if we consider non-homologous loci for use in supertree reconstruction, let alone the current debate about whether and how to infer supertrees.) However, this may not be true for all taxonomic groups. **Since there most likely is a trade-off between the number of taxa and the topological resolution available as input for ParaFit, the user has to decide whether a certain ParaFit analysis based on taxonomic data is worth conducting or not.** It is therefore necessary to closely examine the number of resolved associations (compared to those of a study of the same taxonomic group but based on character data) as well as the resolution of the host or parasite taxonomy trees, as described below.

Computation of the broken-stick distribution (BSD) for the set of associations

The user selects an AT, which results in a new window displaying the BSD (see figure 4). It basically consists of two parts: the first part shows the BSD for the parasites (indicated by a "P" in the first column of each line), the second part the BSD for the hosts (a "H" as identifier). The further columns show the rank of the taxon, its name, its absolute and relative frequency of occurrence within the associations, and its expected frequency according to the BSD. The last column shows whether the real relative frequency is larger than the expected one.

By holding the CTRL button and using the left mouse button, the user can highlight multiple entries within the list. Each entry represents a parasite or host, which is then marked for removal from the association table. Often the entries in the list are highlighted by means of an alternating pattern of dark- and light-grey. Several lines sharing the same colour-scheme correspond to the same tied rank and therefore have to be treated equally.



type	rank	taxon	frequency	proportion	expected	prp. > exp.
P	1	Ustilago_striiformis_307781	54	0.07077	0.02879	yes
P	2	Schizonella_melanogramma_63386	35	0.04587	0.02392	yes
P	3	Microbotryum_violaceum_5272	34	0.04456	0.02148	yes
P	4	Jamesdicksonia_dactylidis_63287	27	0.03539	0.01985	yes
P	5	Ustilago_bullata_117172	21	0.02752	0.01863	yes
P	6	Urocystis_ranunculi_63394	21	0.02752	0.01766	yes
P	7	Thecaphora_saponariae_72562	19	0.0249	0.01684	yes
P	8	Tranzscheliella_hypodytes_349358	18	0.02359	0.01615	yes
P	9	Entyloma_hieracii_189602	18	0.02359	0.01554	yes
P	10	Entyloma_ranunculi-repentis_189607	15	0.01966	0.01499	yes
P	11	Entyloma_microsporum_62642	13	0.01704	0.01451	yes
P	12	Ustilago_avenae_120650	12	0.01573	0.01406	yes
P	13	Ustilago_tritici_117174	11	0.01442	0.01366	yes
P	14	Stegocintractia_luzulae_86812	10	0.01311	0.01328	no
P	15	Tilletia_caries_13290	10	0.01311	0.01293	yes
P	16	Anthracoidea_karii_265844	9	0.0118	0.01261	no
P	17	Tilletia_laevis_157183	9	0.0118	0.0123	no
P	18	Ustilago_hordei_120017	9	0.0118	0.01202	no
P	19	Entorrhiza_casparyana_63375	9	0.0118	0.01174	yes
P	20	Microbotryum_stellariae_288790	8	0.01048	0.01149	no
P	21	Tilletia_bromi_157179	8	0.01048	0.01124	no
P	22	Microbotryum_dianthorum_72560	8	0.01048	0.01101	no
P	23	Ustilentyloma_brefeldii_355612	8	0.01048	0.01079	no
P	24	Anthracoidea_subinclusa_265861	8	0.01048	0.01058	no
P	25	Microbotryum_duriaeanum_162322	7	0.00917	0.01037	no
P	26	Anthracoidea_fischeri_251615	7	0.00917	0.01018	no
P	27	Microbotryum_scorzonerae_72746	6	0.00786	0.00999	no
P	28	Anthracoidea_sempervirentis_265850	6	0.00786	0.00981	no
P	29	Entyloma_fuscum_189600	6	0.00786	0.00964	no
P	30	Tilletia_fusca_157181	6	0.00786	0.00947	no
P	31	Entyloma_fergussonii_318817	6	0.00786	0.00931	no
P	32	Sporisorium_neglectum_307753	6	0.00786	0.00915	no

Figure 4: CopyCat's representation of the Broken-Stick Distribution

N.B.: The broken stick distribution (e.g., Legendre and Legendre 1998, p. 244) is a standard null model of community structure in ecology. It can be used to predict species' relative abundances but may also be used with other kind of data such as, e.g., eigenvectors (Legendre and Legendre 1998, p. 410). Species the relative frequency of which is larger than the corresponding broken stick value occur more frequently than expected by chance.

We have included the BSD here since it may be used to detect host or parasite species which are represented in significantly more associations than others. This is not to say that ParaFit is unable to deal with widespread parasites; on the contrary, these are treated more consistently in ParaFit than in other

cophylogeny programs (Legendre et al. 2002). However, a list of associations derived from literature data may, for instance, include many more associations from host species which are medically or economically important and, thus, have been studied more intensively than their less important relatives. **If the BSD detects species which are represented in a particularly large number of associations, the user may wish to conduct ParaFit runs both before and after exclusion of these taxa.** In case such taxa display a cophylogenetic behaviour strongly deviating from that of other taxa (i.e., significant vs. insignificant associations, or vice versa), presence or absence of these highly frequent taxa may considerably influence global significance. Even though ParaFit's results will (according to Legendre et al. 2002) always be the correct ones given the correctness of the associations, the user may be interested in the impact of such extremely widespread parasites. The broken-stick method provides an objective means to distinguish these from "normal" species.

Filtering of an association table

This step allows the selection of divisions (as defined within the NCBI taxonomy), whose corresponding taxa and, hence, associations should be completely removed from the table.

The "additional filter option" provides a so-called "rank mapping" feature. Each taxon is mapped to its respective parent taxon until the specified rank (e.g. genus or family) is obtained. Redundant entries are removed automatically.

The field "remove parasite/host associations containing specific parasite/host taxon-IDs" allows the removal of associations, whose taxa are listed in this box. Valid input is a list of space-separated taxa IDs – each of them with the leading letter "P" or "H" – indicating the taxon's membership in the group of parasite or host taxa (the taxa being selected via the BSD option are listed in this field).

For example, "H9606" would result in a removal of those associations having a host associated with the taxonomy ID "9606". "Homo_sapiens_9606" would also be a valid input, although only the ID is of interest.

CopyCat - The second tab

This tab (see figure 5) deals with the selection of the input files and parameters for ParaFit (Legendre et al. 2002) and with the ParaFit run itself. Here, the most important feature is a wrapper for the program ParaFit preparing and finally providing the properly formatted input data.

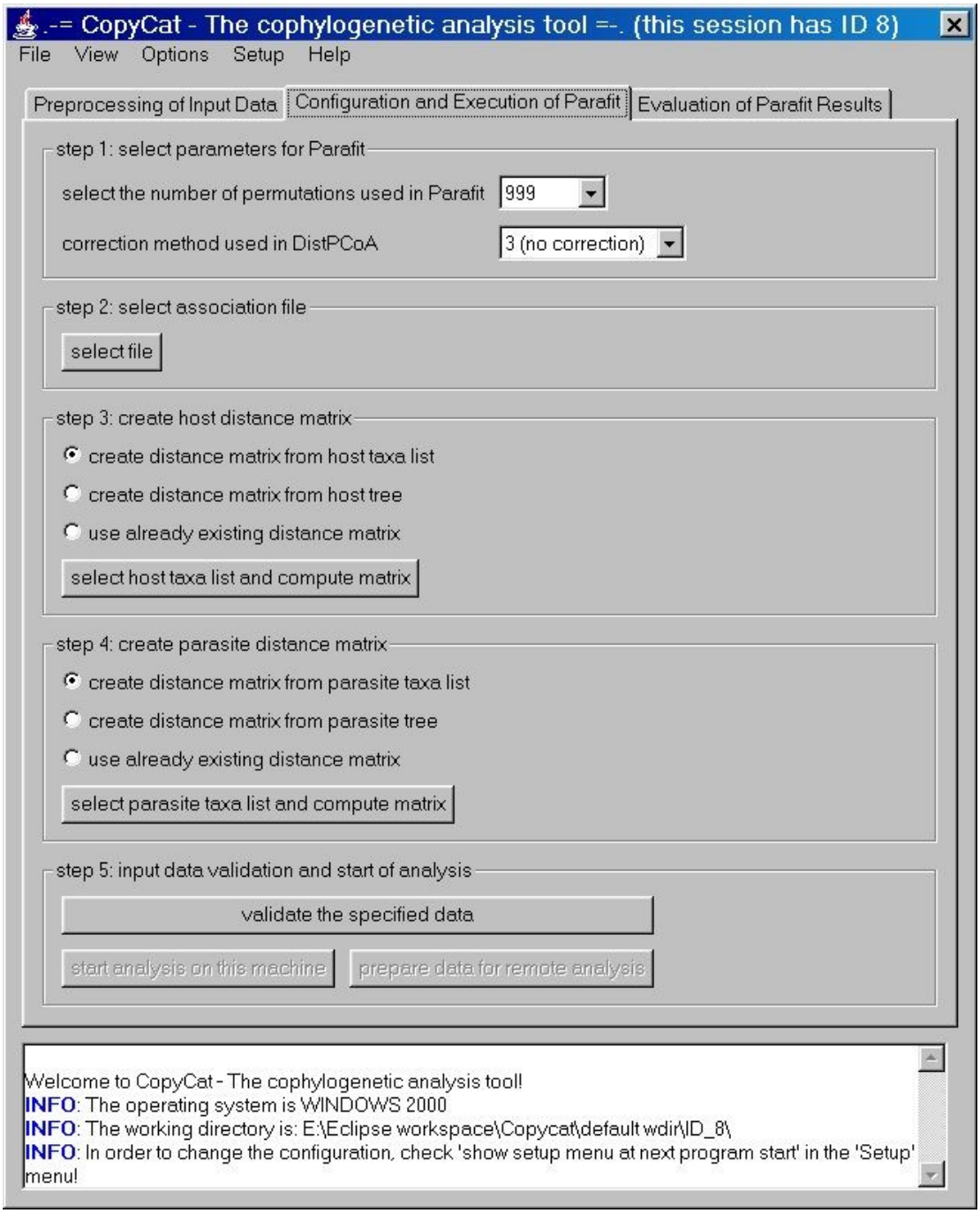

.-= CopyCat - The cophylogenetic analysis tool =-. (this session has ID 8)
File View Options Setup Help
Preprocessing of Input Data
Configuration and Execution of Parafit
Evaluation of Parafit Results
step 1: select parameters for Parafit
select the number of permutations used in Parafit
999
correction method used in DistPCoA
3 (no correction)
step 2: select association file
select file
step 3: create host distance matrix
create distance matrix from host taxa list
create distance matrix from host tree
use already existing distance matrix
select host taxa list and compute matrix
step 4: create parasite distance matrix
create distance matrix from parasite taxa list
create distance matrix from parasite tree
use already existing distance matrix
select parasite taxa list and compute matrix
step 5: input data validation and start of analysis
validate the specified data
start analysis on this machine
prepare data for remote analysis
Welcome to CopyCat - The cophylogenetic analysis tool!
INFO: The operating system is WINDOWS 2000
INFO: The working directory is: E:\Eclipse workspace\Copycat\default wdir\ID_8\
INFO: In order to change the configuration, check 'show setup menu at next program start' in the 'Setup' menu!

Figure 5: The second tab of CopyCat

This tab is divided into 4 steps.

Step 1

The number of permutations per row of the association matrix to be conducted by ParaFit can be specified. ParaFit requires principal coordinates inferred from host and parasite distance matrices as computed by, e.g., the program DistPCoA (Legendre and Anderson 1998b). DistPCoA supports two correction methods for negative eigenvalues: Lingoes method and Cailliez method. The user might choose one of these or simply select the "no correction method" option. The advantages and disadvantages of the several corrections methods are explained in, e.g., Legendre and Anderson (1998a)

and Legendre and Legendre (1998, pp. 432ff.).

Step 2

The user selects an association file.

Step 3 (create host distance matrix)

The user has three possibilities:

- The user selects a pre-existing host distance matrix.
- The user specifies a list of host taxa, which is used to reconstruct the NCBI host tree. The tree is automatically converted to a patristic distance matrix (Farris 1967).
- The user specifies a host tree, which is used to create the distance matrix.

Once a tree is constructed, several of its features are reported to the message window. "Balance" is a measure of tree balance as described by Colless (1982). Note that "balance" can only be computed for rooted binary trees. "Cherry" is the measure of tree balance suggested by McKenzie and Steel (2000) divided by the maximum possible number of cherries (n/2, provided that n is the number of taxa). "Resolution" (described as "Colless' consensus fork index" in Swofford 1991) as well as "information content" (Thorley and Page 2000) are measures of topological resolution. "Resolution" is just the number of internal nodes divided by the maximum possible number of internal nodes (i.e., n-2) and, thus, bound between 0 and 1. A value of 1 indicates full resolution. The cladistic information content has some theoretical advantages, but in case of not fully resolved topologies this measure may rapidly converge towards 0 if the number of taxa increases and, thus, may not be applicable when dealing with large datasets.

N.B.: **Due to the presumed trade-off between the number of taxa and the topological resolution available as input for ParaFit, the user has to base her decision whether to conduct a ParaFit analysis based on taxonomic data not only on the number of resolved associations, but also on the amount of topological resolution.** Even though one of the advantages of ParaFit is that it does not require fully resolved (binary) trees (Legendre et al. 2002), trees may well display not enough topological resolution (too many polytomies) to be of value in conducting cophylogenetic analyses. As an extreme example, consider a totally unresolved taxonomic tree. Since in that case the eigenvectors of all taxa as output by DistPCoA will be identical, such a tree will lead to all associations being insignificant just for trivial reasons.

Step 4 (create the parasite distance matrix)

This is the same as above.

Step 5 (validation)

If all input files have been specified and all parameters been set, the user should then hit the "validate the specified data" button. During this validation,

the program reports whether the taxon names in the association matrix are consistent with the taxon names in the host and parasite distance matrix or not. If taxon names are present in the association table, which are not contained in the respective distance matrix, the program returns an error. In the opposite case (taxon names from the distance matrix can't be found in the association table) the program offers a "shrink distance matrix"-option, which allows the removal of the respective columns and rows from the distance matrix.

N.B.: **In case host or parasite trees are derived by pruning from larger phylogenies, it is much more convenient to change just the association table than to manipulate the trees themselves.** This feature of CopyCat results in a great gain in user flexibility with respect to running ParaFit with slightly different sets of taxa.

If the validation returns no errors, the following two options are enabled:

- "start local analysis": The ParaFit wrapper is started with the parameters specified above.
- "prepare data for remote analysis": As an alternative, all input files, the ParaFit wrapper (including ParaFit and DistPCoA) and a setup file are put into a ZIP archive. The archive can be transferred to a high-performance machine. After archive extraction the wrapper can be invoked by the command "java -Xmx512M -jar ParafitWrapper.jar". The "-Xmx" switch denotes the maximum amount of memory the wrapper has for its own disposal (here: 512 MB).

CopyCat - The third tab

This tab (see figure 6) deals with the evaluation of the ParaFit results.

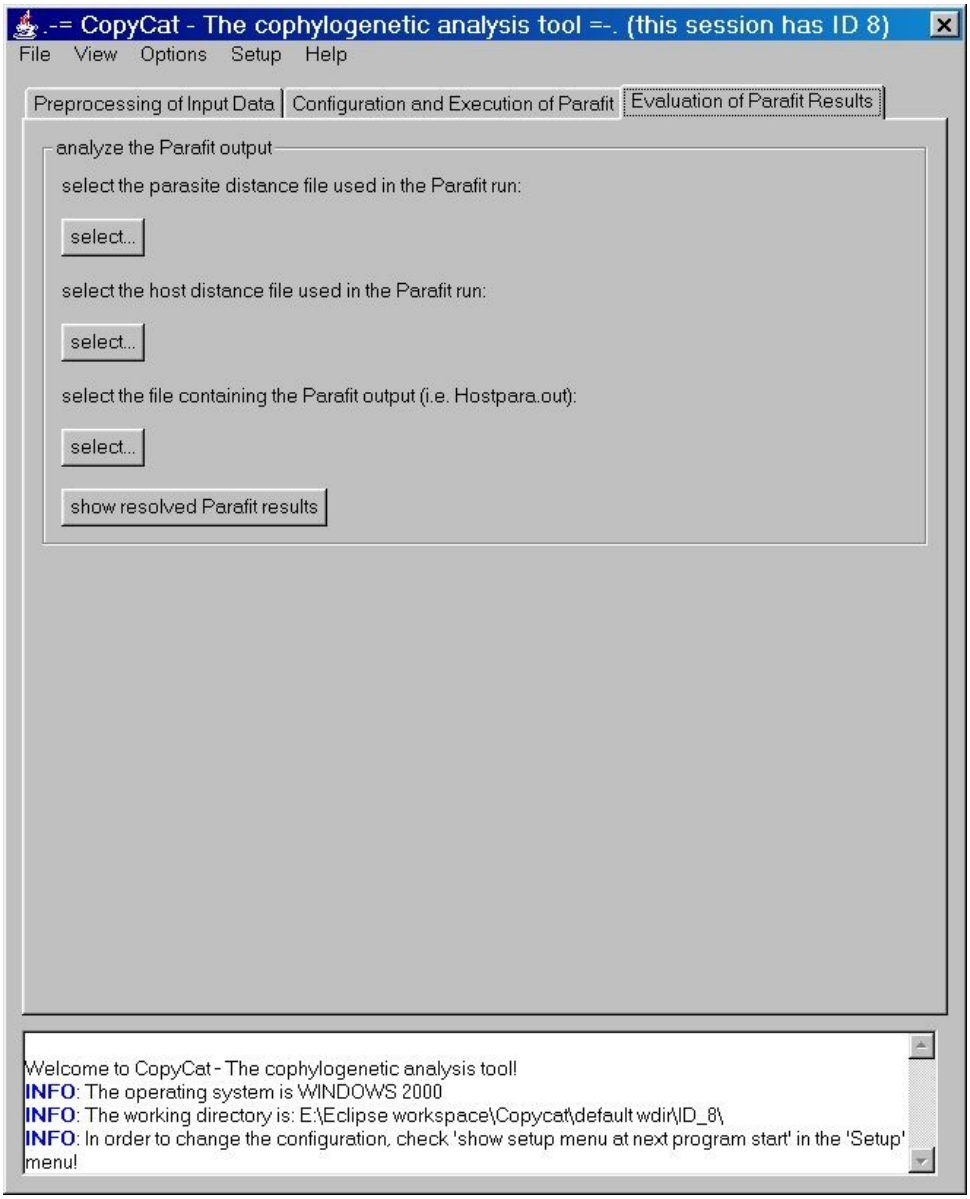

.-= CopyCat - The cophylogenetic analysis tool =-. (this session has ID 8)
File View Options Setup Help
Preprocessing of Input Data
Configuration and Execution of Parafit
Evaluation of Parafit Results
analyze the Parafit output
select the parasite distance file used in the Parafit run:
select...
select the host distance file used in the Parafit run:
select...
select the file containing the Parafit output (i.e. Hostpara.out):
select...
show resolved Parafit results
Welcome to CopyCat - The cophylogenetic analysis tool!
INFO: The operating system is WINDOWS 2000
INFO: The working directory is: E:\Eclipse workspace\Copycat\default wdir\ID_8\
INFO: In order to change the configuration, check 'show setup menu at next program start' in the 'Setup' menu!

Figure 6: The third tab of CopyCat

After the analysis has ended, an output file – called Parafit.out – should have been created. In this step, this output file is specified via the "open"-dialogue, together with the host and parasite distance matrix used in that ParaFit run. The distance matrix files are needed to display the correct organism names, instead of the non-interpretable labels like "Parasite 4" or "Host 17". A sample ParaFit output as resolved by CopyCat is shown in figure 7.

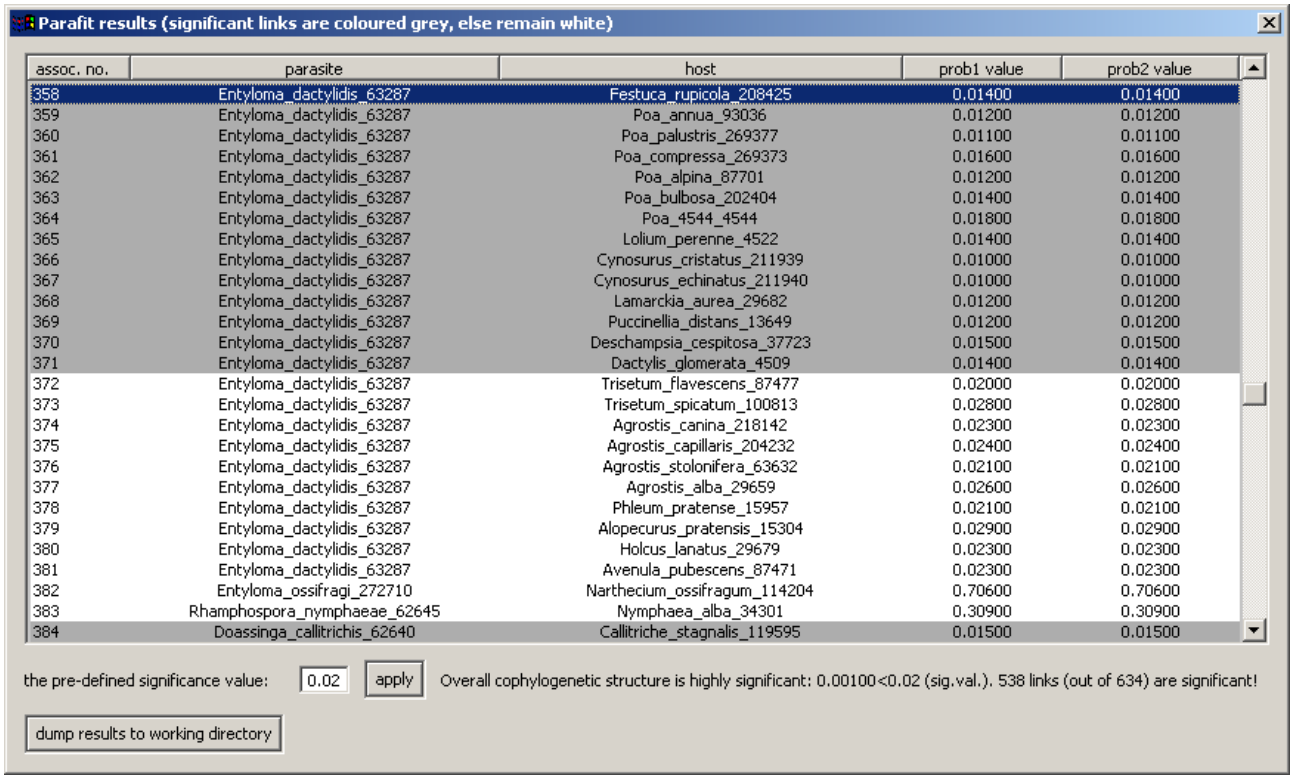

Parafit results (significant links are coloured grey, else remain white)
assoc. no. | parasite | host | prob1 value | prob2 value
358 | Entyloma_dactylidis_63287 | Festuca_rupicola_208425 | 0.01400 | 0.01400
359 | Entyloma_dactylidis_63287 | Poa_annua_93036 | 0.01200 | 0.01200
360 | Entyloma_dactylidis_63287 | Poa_palustris_269377 | 0.01100 | 0.01100
361 | Entyloma_dactylidis_63287 | Poa_compressa_269373 | 0.01600 | 0.01600
362 | Entyloma_dactylidis_63287 | Poa_alpina_87701 | 0.01200 | 0.01200
363 | Entyloma_dactylidis_63287 | Poa_bulbosa_202404 | 0.01400 | 0.01400
364 | Entyloma_dactylidis_63287 | Poa_4544_4544 | 0.01800 | 0.01800
365 | Entyloma_dactylidis_63287 | Lolium_perenne_4522 | 0.01400 | 0.01400
366 | Entyloma_dactylidis_63287 | Cynosurus_cristatus_211939 | 0.01000 | 0.01000
367 | Entyloma_dactylidis_63287 | Cynosurus_echinatus_211940 | 0.01000 | 0.01000
368 | Entyloma_dactylidis_63287 | Lamarckia_aurea_29682 | 0.01200 | 0.01200
369 | Entyloma_dactylidis_63287 | Puccinellia_distans_13649 | 0.01200 | 0.01200
370 | Entyloma_dactylidis_63287 | Deschampsia_cespitosa_37723 | 0.01500 | 0.01500
371 | Entyloma_dactylidis_63287 | Dactylis_glomerata_4509 | 0.01400 | 0.01400
372 | Entyloma_dactylidis_63287 | Trisetum_flavescens_87477 | 0.02000 | 0.02000
373 | Entyloma_dactylidis_63287 | Trisetum_spicatum_100813 | 0.02800 | 0.02800
374 | Entyloma_dactylidis_63287 | Agrostis_canina_218142 | 0.02300 | 0.02300
375 | Entyloma_dactylidis_63287 | Agrostis_capillaris_204232 | 0.02400 | 0.02400
376 | Entyloma_dactylidis_63287 | Agrostis_stolonifera_63632 | 0.02100 | 0.02100
377 | Entyloma_dactylidis_63287 | Agrostis_alba_29659 | 0.02600 | 0.02600
378 | Entyloma_dactylidis_63287 | Phleum_pratense_15957 | 0.02100 | 0.02100
379 | Entyloma_dactylidis_63287 | Alopecurus_pratensis_15304 | 0.02900 | 0.02900
380 | Entyloma_dactylidis_63287 | Holcus_lanatus_29679 | 0.02300 | 0.02300
381 | Entyloma_dactylidis_63287 | Avenula_pubescens_87471 | 0.02300 | 0.02300
382 | Entyloma_ossifragi_272710 | Narthecium_ossifragum_114204 | 0.70600 | 0.70600
383 | Rhamphospora_nymphaeae_62645 | Nymphaea_alba_34301 | 0.30900 | 0.30900
384 | Doassinga_callitrichis_62640 | Callitriche_stagnalis_119595 | 0.01500 | 0.01500
the pre-defined significance value:
0.02
apply
Overall cophylogenetic structure is highly significant: 0.00100<0.02 (sig.val.). 538 links (out of 634) are significant!
dump results to working directory

Figure 7: CopyCat's representation of the ParaFit output. By selecting "dump information to working directory" the results are stored in a simple text file (ASCII format). Lines contained in that file marked by a "+" character represent significant links, while lines starting with a "-" character are considered as "non significant".

CopyCat – Available menu bar options

The menu bar offers the following options:

File

"Transfer content of working directory to a place of your choice"

Once your work with CopyCat is done, you might want to transfer data from the working directory to a directory of your choice. This can be done by selecting this option.

"Download NCBI taxonomy file(s)"

As mentioned above, it is advisable to get the latest NCBI taxonomy file from time to time. The taxonomy itself is steadily improved due to the incorporation of more recent phylogenetic insights, and the total number of both terminal taxa and taxa of higher rank included in the taxonomy dump files is steadily increasing, potentially increasing both the number of resolved associations and the topological resolution in the NCBI-based cophylogenetic analyses. By checking this option, CopyCat downloads the latest NCBI taxonomy dump file and places it in the appropriate directory.

View

"View content of working directory"

This shows a view of the current working directory and its content.

Options

"Enable Strict Filtering of Association Table"

By default the filter process scans an association table and removes all associations (=lines), which do not fulfill one of the following criteria. Both, parasite and host label, ...

- ... have to exist in the NCBI taxonomy
- ... should not be blacklisted.

If the user has provided an association table containing custom taxon labels (such as „Patient234“ instead of „Homo sapiens“), the program would remove this line from the new associations table due to the first condition („... have to exist in the NCBI taxonomy ...“). Even though, this condition can be relaxed by **unchecking** the option „ Enable Strict Filtering of Association Table“ in the menu bar.

"Use Equal Branch Length (=1) for tree2dist Conversion"

If you have specified a tree file in CopyCat's second tab, you might want to have topological distances in the distance matrix resulting from that tree. By checking this option branch lengths are set to 1. By default this option is not checked. In that case, a patristic distance matrix results in which each pairwise distance between two taxa A and B represents the sum of the number of taxa in which A is contained (including A), but not B, and the number of taxa in which B is contained (including B), but not A. Accordingly, by default branch lengths in the NCBI taxonomy tree may be larger than one, representing more information extracted from the taxonomy than just the topology.

"Use AxPcoords and AxParafit instead of DistPCoA and Parafit"

By selecting this option, AxPcoords and AxParafit are used instead of Legendre's programs (Legendre et al. 2002). Originally, CopyCat was designed for supporting the programs DistPCoA (computation of principal coordinates from a distance matrix) and Parafit (providing tests for both overall phylogenetic congruence as well as for the significance of individual host-parasite associations). As of April 2007, CopyCat supports Alexandros Stamatakis' programs AxParafit and AxPcoords, which are highly optimized versions of Parafit and DistPCoA, respectively.

CopyCat searches within the “code” subdirectory, which contains all executables for the analyses. By default, this folder contains the following

operating-system-dependant executables:

- parafit_win.exe
- distpcoa_win.exe
- AxParafit_win.exe
- AxPcoords_win.exe

The attribute “win” denotes the executables for the Windows platform (accordingly, “mac” for Macintosh and “linux” for Linux systems are used). In case the user has selected the above option, CopyCat chooses “AxParafit_win.exe” and “AxPcoords_win.exe” (operating system: Windows). If you wish to use externally compiled executables (e.g., using the ACML[3] or MKL[4] libraries), they should obey to the naming convention <PROGRAM>_<OS>_<LIBRARY> as, e.g.,

- AxParafit_win_MKL.exe
- AxPcoords_win_MKL.exe

and moved into the “code” subdirectory. CopyCat will detect and use them for the next computation. Such an optimized executable will always be preferred to the default one.

Once the user has selected the AxPcoords/AxParafit option, the option “correction method used in DistPCoA” (the second item on the second tab) will be disabled because correction methods for negative eigenvalues are not supported by AxPcoords.

Setup

"Show setup menu at next program start"

If you want to change the working directory, you just have to check this option. On the next start of CopyCat the configuration dialogue will appear again.

Tutorial - A step-by-step example run of Copycat

This tutorial focusses on the kind of input needed for certain steps in Copycat and shows the output produced by that input. The underlying data set for this example run is the list of European smut fungi and their hosts from Vánky ([Ván94],[Ván05]). This data set does not contain parasite or host trees or distance matrices, so we have to construct them using the NCBI taxonomy.

1st) The input: An association table of smut fungi (parasites) and their respective hosts.

Here is an excerpt of the input file 'smut_fungi_association_table.txt':

3 AMD © Core Math Library
4 Intel © Math Kernel Library

```
Anthracoidea altera                 Carex fuliginosa
Anthracoidea angulata               Carex hirta
Anthracoidea arenaria               Carex arenaria
Anthracoidea arenaria               Carex brizoides
Anthracoidea arenaria               Carex ligerica
Anthracoidea arenaria               Carex ovalis
Anthracoidea arenaria               Carex praecox
Anthracoidea aspera                 Carex appropinquata
Anthracoidea aspera                 Carex chordorrhiza
Anthracoidea aspera                 Carex diandra
Anthracoidea aspera                 Carex glareosa
Anthracoidea baldensis              Carex baldensis
        ...                                 ...
```

Text 1: The unresolved association table. Parasites are in the left column - hosts in the right column. Each parasite is separated from its respective host by a tab character. The file contains 1853 host-parasite associations.

2nd) Resolving the association table: Each parasite/host contained in the NCBI Taxonomy Database has a unique taxonomy ID. This step tries to gather these IDs. Each association containing both, a parasite **and** a host with a valid ID, is used in the resulting so-called “resolved association table”. A representation of this resolving step is shown below (see illustration 1).

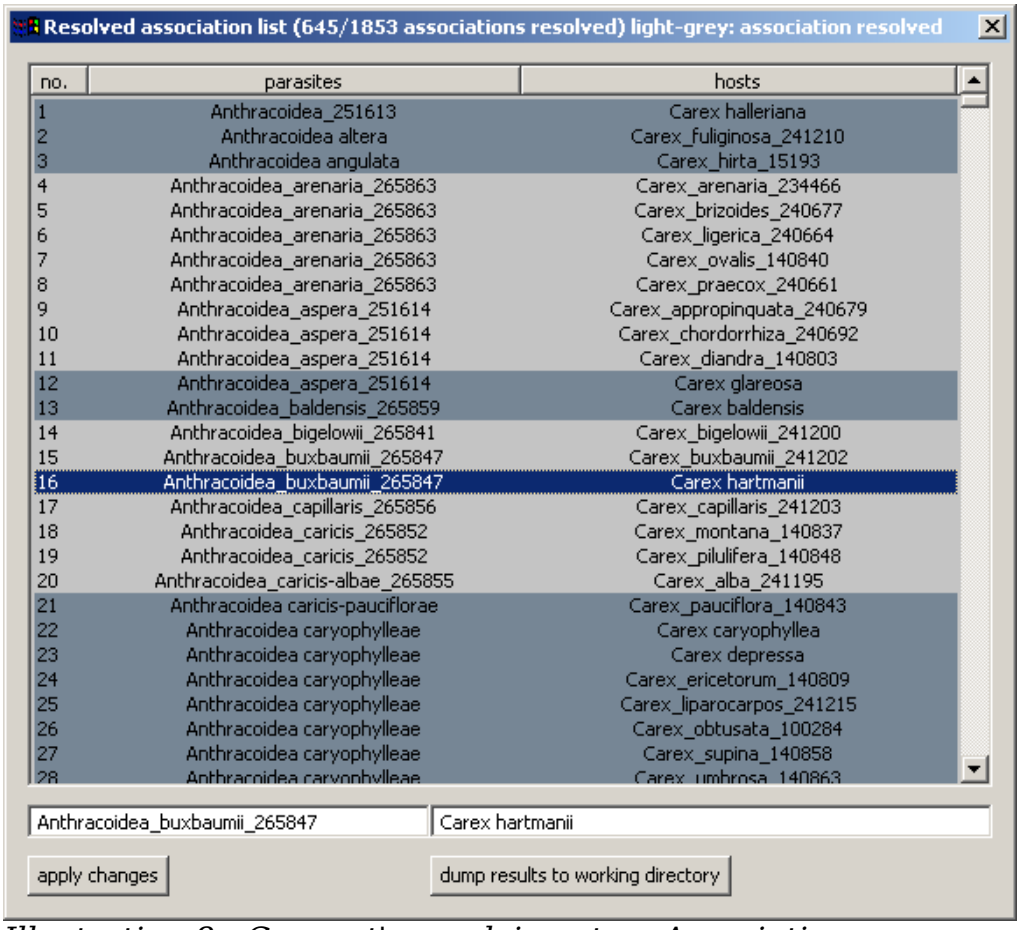


Illustration 8: Copycat's resolving step. Associations highlighted in light-grey have a valid taxonomy ID for both the parasite and the host. By clicking the 'dump results to working directory' button the resolved association table is written to the working directory (a subset of the original table). The file name contains a '_resolved.txt' suffix.

3rd) Generation of the parasite/host taxa lists. For the Parafit [LDB02] analysis of this association data, we first need to draw two lists from the resolved association table: one containing all parasite taxa and another one containing all host taxa.
This is achieved by selecting '*apply settings to association list*'. Here, the specified association list can be filtered in regard to certain criteria and as a side-effect the parasite/host taxa lists are written to the working directory. Naturally, the user is not obliged to select certain filter criteria but can simply choose the '*leave associations in their current state*' option. Consequently, the specified association table stays untouched.
In this tutorial we make use of the latter and issue the association table gained in the previous section. This operation will take a moment. Finally, the following two files appear in the working directory:

'hosts_filtered_using_option_0.txt' 'parasites_filtered_using_option_0.txt'

The '0' indicates that we selected the '*leave associations in their current state*' option.
4th) Creation of a host distance matrix and a parasite distance matrix. The taxa

lists from the previous step are now being used for the creation of the respective NCBI host tree and NCBI parasite tree. Once this is done, the respective distance matrices are generated.
We switch to Copycat's second tab and select the '*distance matrix from host taxa list*' option together with the 'hosts_filtered_using_option_0.txt' file. This results in the call of the *ParafitWrapper*. The wrapper will now try to create the denoted host distance matrix. You might want to follow the process of the wrapper by reading the lines in the message window that are marked by the purple “WRAPPER” tag. The wrapper is finished once the message window contains the following three information lines at the very end (here: example values).

INFO on host tree: Resolution: 0,252525
INFO on host tree: Balance: input is no rooted binary tree
INFO on host tree: Information content: Infinity

The working directory should now contain the files 'hosts.dist' and 'host.out.tree'.
We repeat the procedure for the 'parasites_filtered_using_option_0.txt' file by selecting the '*distance matrix from host taxa list*' option in the box below. The distance matrix is finished once the following three lines appear at the end of the message window (here: example values).

INFO on parasite tree: Resolution: 0,252525
INFO on parasite tree: Balance: input is no rooted binary tree
INFO on parasite tree: Information content: Infinity

The working directory should now contain the files 'parasites.dist' and 'parasites.out.tree'. These info messages show the resolution and the phylogenetic information content (which can't be computed for large numbers of taxa) of the host/parasite tree created. If the resolution is not satisfying enough you have the option to cancel the further cophylogenetic analysis at this stage.

If you want to proceed, you have to choose the '*validate button*'.
The validation step ensures that the taxa contained in the following three files are consistent with respect to each other. This means that a parasite's name contained in the association table should exist in the parasite distance matrix and so forth.

If the validation is successful, you will have the option to immediately start the Parafit run on this machine or, as an alternative, to pack all relevant files to a zip archive. This archive can be transferred to another, probably more powerful machine.

Once the *ParafitWrapper* has started (either on this or another machine) it generates a random session ID (similar to Copycat) and creates a respective subdirectory within your working directory. This directory is named after that session ID. The time the wrapper is running any interaction with Copycat is blocked.
Depending on the size of the input data, the Parafit run can be a time-consuming issue.

5th) Analysis of the Parafit results. Parafit performs tests for both the overall phylogenetic congruence as well as the significance of individual associations.

These results are listed in the file 'HostPara.out'. Copycat reports the location of 'HostPara.out' by printing:

WRAPPER: Please, check file [D:\Copycat\defaultwdir\ID_3738\ID_8313\Hostpara.out] for results!

'HostPara.out' holds the results on the individual links in the following format:

Parasite 17 **Host**154 **F1** =******* **Prob1** = 0.94600 **F2** = -0.00016 **Prob2** = 0.99200

As 'Parasite 17' and 'Host 154' provide no information on the actual organisms, Copycat needs to know the location of the parasite and the host distance file used in that specific Parafit run. These files should reside in your working directory.
Once the three files ('Hostpara.out', parasite distance matrix and host distance matrix) have been specified, Copycat opens a new window showing the results. An example is shown in illustration 2.

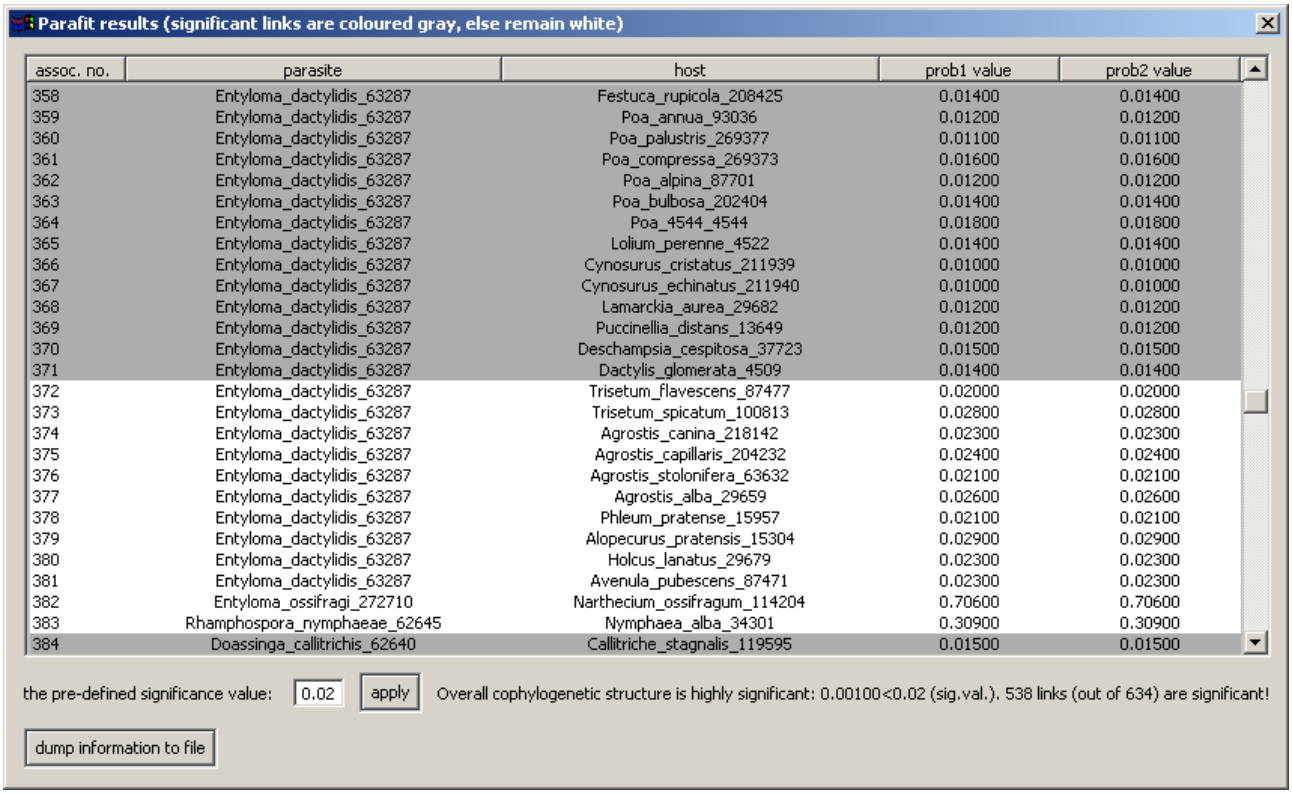

Parafit results (significant links are coloured gray, else remain white)
assoc. no. | parasite | host | prob1 value | prob2 value
358 | Entyloma_dactylidis_63287 | Festuca_rupicola_208425 | 0.01400 | 0.01400
359 | Entyloma_dactylidis_63287 | Poa_annua_93036 | 0.01200 | 0.01200
360 | Entyloma_dactylidis_63287 | Poa_palustris_269377 | 0.01100 | 0.01100
361 | Entyloma_dactylidis_63287 | Poa_compressa_269373 | 0.01600 | 0.01600
362 | Entyloma_dactylidis_63287 | Poa_alpina_87701 | 0.01200 | 0.01200
363 | Entyloma_dactylidis_63287 | Poa_bulbosa_202404 | 0.01400 | 0.01400
364 | Entyloma_dactylidis_63287 | Poa_4544_4544 | 0.01800 | 0.01800
365 | Entyloma_dactylidis_63287 | Lolium_perenne_4522 | 0.01400 | 0.01400
366 | Entyloma_dactylidis_63287 | Cynosurus_cristatus_211939 | 0.01000 | 0.01000
367 | Entyloma_dactylidis_63287 | Cynosurus_echinatus_211940 | 0.01000 | 0.01000
368 | Entyloma_dactylidis_63287 | Lamarckia_aurea_29682 | 0.01200 | 0.01200
369 | Entyloma_dactylidis_63287 | Puccinellia_distans_13649 | 0.01200 | 0.01200
370 | Entyloma_dactylidis_63287 | Deschampsia_cespitosa_37723 | 0.01500 | 0.01500
371 | Entyloma_dactylidis_63287 | Dactylis_glomerata_4509 | 0.01400 | 0.01400
372 | Entyloma_dactylidis_63287 | Trisetum_flavescens_87477 | 0.02000 | 0.02000
373 | Entyloma_dactylidis_63287 | Trisetum_spicatum_100813 | 0.02800 | 0.02800
374 | Entyloma_dactylidis_63287 | Agrostis_canina_218142 | 0.02300 | 0.02300
375 | Entyloma_dactylidis_63287 | Agrostis_capillaris_204232 | 0.02400 | 0.02400
376 | Entyloma_dactylidis_63287 | Agrostis_stolonifera_63632 | 0.02100 | 0.02100
377 | Entyloma_dactylidis_63287 | Agrostis_alba_29659 | 0.02600 | 0.02600
378 | Entyloma_dactylidis_63287 | Phleum_pratense_15957 | 0.02100 | 0.02100
379 | Entyloma_dactylidis_63287 | Alopecurus_pratensis_15304 | 0.02900 | 0.02900
380 | Entyloma_dactylidis_63287 | Holcus_lanatus_29679 | 0.02300 | 0.02300
381 | Entyloma_dactylidis_63287 | Avenula_pubescens_87471 | 0.02300 | 0.02300
382 | Entyloma_ossifragi_272710 | Narthecium_ossifragum_114204 | 0.70600 | 0.70600
383 | Rhamphospora_nymphaeae_62645 | Nymphaea_alba_34301 | 0.30900 | 0.30900
384 | Doassinga_callitrichis_62640 | Callitriche_stagnalis_119595 | 0.01500 | 0.01500
the pre-defined significance value:
0.02
apply
Overall cophylogenetic structure is highly significant: 0.00100<0.02 (sig.val.). 538 links (out of 634) are significant!
dump information to file

Illustration 9: Copycat's representation of the Parafit results

References

Colless, D. H. (1982). Review of Phylogenetics: The Theory and Practice of Phylogenetic Systematics, by E.O.Wiley. Syst. Zool. 31: 100-104.

Farris, J.S. 1967. The meaning of relationship and taxonomic procedure. Systematic Zoology 16: 44–51.

Legendre, P. and Anderson, M. J. 1998a. Distance-based redundancy analysis: testing multi-species responses in multi-factorial ecological experiments. Ecological Monographs 69: 1-24.

Legendre, P. and Anderson, M. J. 1998b. Program DistPCoA. Departement de sciences biologiques, Universite de Montreal. 10 pages. Available at http://www.bio.umontreal.ca/Casgrain/en/labo/distpcoa.html

Legendre, P. and Legendre, L. 1998. Numerical Ecology. Second English Edition. Elsevier, Amsterdam, 853+XV pages.

Legendre, P., Desdevises, Y., and Bazin, E. 2002. A statistical test for host-parasite coevolution. Systematic Biology 51(2): 217-234.

McKenzie, A. and Steel, M. 2000. Distributions of cherries for two models of trees. Math. Biosci. 164: 81–92.

Swofford, D. 1991. When are phylogeny estimates from molecular and morphological data incongruent? Pp. 295–333 in: Miyamoto, M.M. and Cracraft, J. (Eds.), Phylogenetic analysis of DNA sequences. Oxford University Press, New York/Oxford.

Thorley, J.L. and Page. R.D. 2000. RadCon: phylogenetic tree comparison and consensus. Bioinformatics, 16(5): 486–487.

Vánky, K. 1994. European smut fungi. Gustav Fischer Verlag, Stuttgart, Jena, New York.

Vánky, K. 2005. European smut fungi (Ustilaginomycetes p.p. and Microbotryales) according to recent nomenclature. *Mycologia Balcanica*, **2**:169-177, 2005.